Data protection compliance challenges for self-sovereign identity

Alexandra Giannopoulou[1][000-0002-7692-8062]

[1] Institute for Information Law (IViR)- University of Amsterdam, The Netherlands
a.giannopoulou@uva.nl

Abstract. Various identity management solutions are emerging in different jurisdictions, with the goal of creating a unified and privacy-preserving identity management system bridging the offline with the online. Within this trend, the concept of self-sovereign identity has re-emerged. It is a concept attached to expressions of both individual autonomy and individual control (sovereignty) — an aspiration in direct relation to what blockchain is promised to bring in contemporary discourse. The paper will provide an overview of the current self-sovereign identity paradigm solutions within the technological environment that involves decentralized networks and it will trace some of the challenges it faces within the European Union especially with regards to the General Data Protection Regulation (EU) 2016/679 (GDPR).

Keywords: self-sovereign identity, GDPR, data protection.

1 Introduction

Humans are just the sort of organisms that interpret and modify their agency through their conception of themselves.
This is a complicated biological fact about us.
Amelie Rorty

Identity management through decentralized ledger applications has been on the forefront of the technological innovation agendas of public institutions, private companies, and privacy-aware communities. The paper will provide an overview of the current self-sovereign identity paradigm solutions and the challenges it faces within the European Union especially with regards to the General Data Protection Regulation (EU) 2016/679 (GDPR).

There is no uniform rule about what constitutes a person's identity; the concept and its governing norms shift according to the legal, technological or institutional context. In the past decades, the digital expansion of ourselves has shaped the idea of the creation of a "digital identity". This, coupled with the overproduction of personal data in the current data-intensive technological environment that has formed our data-driven societies has created a newly found interest in preserving privacy and data protection

online. Scandals such as Cambridge Analytica have illustrated that there are significant shortcomings in the current data management and data governance mechanisms. Hence, the social and legal circumstance have opened up the potential for envisaging a technological version of a digital identity that solves current insufficiencies.

The creation of a new and uniform digital identity ecosystem is an aspiration that has progressively risen into prominence in disperse ways. Various identity management solutions are emerging in different jurisdictions, with the goal of creating a unified -privacy-preserving- identity bridging the offline with the online. The market of digital identity is already quite substantial and very diverse. It aims to provide a technological solution to financial inclusion, reputation management, privacy-preserving social media identities etc. Within this trend, the concept of self-sovereign identity has re-emerged[1]. No consistent definition of the concept has been established. In general terms, we can describe self-sovereign identity as *an identity management system, developed by a private or public entity which takes technological design decisions for personal data management guided by a set of principles that are loosely defined and not universally accepted as a common standard.* It is essentially a technological solution which transcribes the goal of autonomy and individual control through decentralization and "user-centric design" over the usage, storage and transfer of one's digital data.

The concept is attached to expressions of both individual control (sovereignty) and trusted verifiability— an aspiration familiar to what blockchain is promised to bring in contemporary data protection discourse [1]. The identity management solutions are several and they all rely on the use of decentralized ledgers, cryptography, and local processing of data. The application of these technological design options aims to materialize some of the core principles of lack of central authority that controls the identity data, of decentralized verifiability and privacy. Due to the granular prioritization of the purported design principles, and the progressive distancing of current technological state of the art from the self-sovereign ideological underpinnings, "decentralized identities" is being used interchangeable although still suffering from the same semantic uncertainties. With coinciding objectives and features, self-sovereign identity projects have become increasingly attached to blockchain technological development and mainstream adoption. At the same time, blockchain enthusiasts are hinging on the success of self-sovereign identity solution as the first implemented use case of blockchain technology.

---

[1] Among the first references to the concept of a self-sovereign identity -as it stands today- can be found here: https://www.moxytongue.com/2012/02/what-is-sovereign-source-authority.html
The author (Devon Loffreto) uses the term self-sovereign authority to describe an identity that does not hinge on external recognition and verifiability, but that it is self-verifiable and controlled by the individual as an entity and not as a citizen. The concept, later revamped as self-sovereign identity, was popularized by cryptographer Cristopher Allen. We use the terms self-sovereign identity and decentralized identity interchangeably.

The expansion of decentralized identity solutions, the growing market, and the institutional interest all bring out questions with regards to the legal framework surrounding their implementation. The eIDAS Regulation[2] defines levels of trust services and provides thus the regulatory environment that enables the creation of different legally compliant identity systems solutions. In addition, the compliance challenges depend on ensuring GDPR compliance and on the establishment of accountability mechanisms within the actors involved. Finally, the applicable legal norms are domain-dependent, with certain areas being highly regulated (i.e. financial markets and institutions). These conciliations are at times harder to achieve

The paper will focus on the compliance challenges that self-sovereign identity solutions face when operating within the scope of territorial application of the GDPR and the obligations that this entails. On the one hand, we discuss the shared vision that blockchain technologies and self-sovereign identity have and on the other, we present some of the key compliance mechanisms that self-sovereign identity solutions will need to address in order to offer a product that delivers on its promises. We conclude with suggestions on conciliations between the data protection regulatory framework and then GDPR.

2 The shared vision of control

Visions of a self-sovereign self, have been attached to different political ideologies. It was only in 2016 that the fundamental design principles of the concept of self-sovereign identity came to life in a form of check list of design options by cryptographer Christopher Allen. In his blog[3], he describes the core principles for the creation of an identity ecosystem that is controlled by each individual and does not hinge on a specific powerful technological infrastructure nor a private or public entity. After tracing the evolution of identity management systems -from centralized to federated to user-centric- the author points out that the self-sovereign identity goes a step beyond the previous systems in that it prioritizes user autonomy through ten foundational principles: existence, control, access, transparency, persistence, portability, interoperability, consent, minimalization, and protection. These principles ensure that the user remains the sole gatekeeper of their respective personal data that constitute the identity that actors will seek to user in order to provide respective services. The concept is goal oriented, focused on preserving "the right for the selective disclosure of different aspects of one's identity and the various components thereof, in different domains and contextual settings"[4].

[2] Regulation (EU) N°910/2014 on electronic identification and trust services for electronic transactions in the internal market adopted on 23 July 2014.

[3] Allen, C (2016), The Path to Self-Sovereign Identity, 25 April 2016, Life With Alacrity blog, http://www.lifewithalacrity.com/2016/04/the-path-to-self-soverereign-identity.html

From a legal standpoint, the General Data Protection Regulation[4] was designed to provide the legal framework with the appropriate assurances that enable individual control over personal data. As a matter of fact, recital 7 of the GDPR directly highlights that "natural persons should have control of their own personal data". This control principle is conveyed through a set of accountability measures that are imposed on responsible actors and a set of rights assigned to each data subject in order to empower them to exercise control over their personal data. The technological solution of the decentralized identity relies heavily on informed consent, as the regulatory representation of the expression of the autonomy of natural persons. [5]

From a technological standpoint, decentralization implies that "no trusted third party should be given control of data, but instead individuals and groups should maintain control over their own data" [2] Distributed ledger technology is a priori embodying the principles of control, security, and transparency with one of is most prominent promises within the current market being "data sovereignty". According to Sullivan, blockchains could enable individuals "to control access to their identity information and to create, manage and use a self-sovereign identity" [3]. As a matter of fact, blockchains provide the technological guarantees for trusted data sharing: they permit modularity, transparency and security though encryption. The techno-legal and ideological circumstance has guided the flourishing of a market that proposes diverse identity solutions on a decentralized environment. Depending on how they are designed, these solutions can convey a granular level of decentralization [6]; not all decentralized identities will be thus considered similar. The variety of the solutions is due not only to the diverse technological design choices but also to the entity that develops them.

With the principles of a decentralized identity being constantly in flux, the standardization of technological architectures constituted a first effort towards harmonizing and establishing a legally compliant and technologically secure set of identity solutions. Among the existing solutions, the W3C has launched a set of standard setting processes for decentralized identity in order to provide a unified strategy towards a common aspiration of eradicating centralized control of personal data. Whether the consortium will succeed in enabling a decentralized identity infrastructure or whether the mistakes of previous standardizing attempts on the Web will be repeated remains still unclear. "Will the blockchain revolution bring a new decentralized web into existence, or simply become the technical infrastructure of further control and centralization?" [2] According to the W3C existing decentralized identity technical documentation [4], "Decentralized identifiers (DIDs) are a new type of identifier to provide verifiable, decentralized digital identity. These new identifiers are designed to enable the controller of a DID to prove control over it and to be implemented independently of any centralized registry, identity provider, or certificate authority". The technical specifications describe the control of public and private keys by the individual (or data subject in legal terms) through the use of a DID document in order to autonomously manage the information related to them.

The DID document is defined as "a set of data describing the DID subject, including mechanisms, such as public keys and pseudonymous biometrics, that the DID subject

[4] Hereinafter GDPR.

can use to authenticate itself and prove their association with the DID. A DID document might also contain other attributes or claims describing the subject". The qualification of these data as personal depends on the data protection test of identifiability which ultimately leads to the qualification of data as anonymous or pseudonymous and thus, personal. Seeing as in the context of the semantic Web digital identifiers can be attributed to a multiplicity of entities such as Internet of Things, companies etc., the GDPR compliance questions refer only to those DIDs that are used to manage data that refer to natural persons.

3 Rights and obligations in a self-sovereign ecosystem

When the decentralized identifiers are created, they can be used in a multiplicity of ways. They can be stored in the personal device of the user, verifiable credentials referring to that individual can be put on the distributed ledger in various privacy-preserving forms, and they can be also transmitted to third parties and entities. Maintaining the ledger of transactions of verifiable credentials necessitates a network of nodes that process and collectively keep the distributed database up to date. The pressing compliance questions in terms of the GDPR relate first to whether the data published in the -public or private- typically permissioned networks can be qualified as personal and if they are indeed personal, what are the accountability obligations among the p2p network of nodes. Depending on the design characteristics of the network, the permissioned network signifies that only the validating participating nodes will have "write" privileges on the ledger. These will be the only ones having privileges of storing, processing and transmitting -personal- data. While these questions have been addressed beforehand for public blockchains in general [1], the specific technological set developed for decentralized identities provides a new breeding ground on which GDPR compliance can be examined.

3.1 Personal data processing

The generation of the decentralized identity starts with the issuance of a verifiable credential stored at the individual's device, which contains the public and private keys belonging to the user. The keys in question constitute part of the identity of the user and are considered pseudonymous data according to the WP29's opinion, which underlines that asymmetric encryption methods are pseudonymisation methods that "merely reduce the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure". Pseudonymous data are protected by the GDPR according to article 4(5) GDPR, as personal data. According to Recital 30 of the GDPR, "natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses (…) or other identifiers (…) This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them." Public keys fall into this category, and their

management falls into the risk assessment obligation of data controllers, per the Article 25 GDPR privacy by design obligations. The French Data Protection Authority (CNIL) has issued an official opinion on blockchains arguing that "the very architecture of blockchains means that these identifiers are always visible, as they are essential for its proper functioning. The CNIL therefore considers that this data cannot be further minimised and that their retention periods are, by essence, in line with the blockchain's duration of existence". Therefore, it can be argued that the public keys, combined with necessary privacy enhancing mechanisms (PETs) could potentially fulfil the data minimisation requirements of the GDPR.

Besides the question of public and private keys, the generation of hashes that serve as attestations of the transaction of verifiable credentials will be put on-chain. In order to respond to the question of whether these hashes fall under the personal data qualification of the GDPR, one has to refer to the notion of risk is pervasive across the Regulation [8]. For example, in article 25 GDPR, the privacy by design obligation is measured through the concept of risk. As a matter of fact, the data controller(s) obligations to implement technical and organizational measures has to be considered according to the case-specific processing that will take place and in order to minimize the risks for the data subjects' rights. The goal of the data protection legislation is not to exclude risk or to ensure that it does not manifest in any form during the data processing. Rather, the legislator embraces the risks involved in personal data processing and employs a wide variety of tools (accountability and obligations to responsible actors, data subjects' rights etc) to minimize the risk involved.

The Regulation encourages pseudonymisation in order to 'reduce the risks to the data subjects concerned'. According to the GDPR, pseudonymisation of data does not equal anonymization. Pseudonymous data are subject to GDPR restrictions. However, the distinction between the two methods is not always clear. The criteria for distinction can be found first on recital 26 GDPR, which specifies that data becomes anonymous if it is 'reasonably likely' that no identification of a natural person can be derived. An individual is considered to be 'identifiable' where they can be 'distinguished' from others. In previous reports, the Article 29 Working Party (which is now the European Data Protection Board) have provided a more absolute interpretation when it comes to various methods of processing of personal data. The two approaches, relative and absolute conflict regularly. The risk-based approach implies that the determination of the risk inherent in the likelihood to re-identify falls on the data controllers. There is tension between the risk-based factor introduced through the GDPR and the absolute approach that existed thus far. The national DPAs' opinions reflect this lack of homogeneity. For example, according to the Irish DPA, the data have to be rendered "irreversibly" anonymous but the criterion of irreversibility is applied in a more relative manner linking it to the absence of reasonable likelihood of identifiability. Similarly, the French Data Protection Authority (CNIL) acknowledges that anonymization tends to make identifiability "practically impossible".

Within this normative framework, the A29WP has published an absolute opinion when it comes to hashing as a method of pseudonymization. A more recent report published by the Spanish Data Protection Authority nuances the absolute approach by introducing the notion of risk in its assessment of the technological method. Hence, and

according to the Spanish DPA, hashing can at times be considered as anonymization or pseudonymization depending on a variety of factors varying from the entities involved to the type of the data at hand.

According to Recital 26 GDPR, the distinction criteria between pseudonymization and anonymization can be found in the "the means reasonably likely to be used (..) either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments." The time frame of the appreciation of these factors for the identifiability test is relative according to the GDPR, in conjunction with the available technology.

The risk assessment in the case of recordation of data on blockchains should take into consideration the envisioned timeframe for the technology at hand. Factors related to the technology, such as the append-only nature of the blockchain, have to be taken into account when appreciating the efficiency of a specific technical anonymization method chosen. Within that frame, taking into consideration organisational measures to employ can be a risk-minimizing factor for the obligations that the data controllers are facing. When the issuer presents the hash of the credential on the blockchain, it is important to assess the likelihood of identification according to the person or entity that would try to identify. This assessment will have to include perspectives of third parties and of the data controllers, and possible de-identification brute forcing methods such as content-based reidentification.

3.2 Actor accountability

According to GDPR, there are two types of actors whose role and relationship within the data protection environment attributes them a set of obligations and responsibilities to abide by data protection rules. Data controllers are responsible according to article 24(1) GDPR to make sure that the data processing is in compliance with the Regulation and to make sure that data subjects have the ability to exercise their rights. This set of obligations is directed at the entities responsible to take "technical and organizational measures" enforcing the GDPR rules. They will be liable to pay compensation for damages ensued in case these measures assume too high a risk towards personal data processing in case of unlawful processing. Data controller is defined in article 4(7) GDPR as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law". Strictly as defined by law, the concept of the data controller requires further clarifications in order to fit a decentralized model of data processing. As a matter of fact, the centralized linear practices that depend on clearly defined boundaries between data subjects, data controllers, and data processors

are constantly challenged and contradicted in today's technological realities. There is a clear need for interpretations, guidelines, and rules that are progressively created by the regulatory framework and its respective case law, as well as the European Data Protection authorities.

A *de facto* analysis of the circumstances of the data processing will determine the entities identified as data controllers, regardless of what is stated in a previous written contractual agreement between the different participating entities and actors. The current interpretation of what it means to define the "means and purposes of the processing" is considered to be broad enough in order to ensure the applicability of these actors on a networked environment. According to the current interpretations of responsibility allocation among data controllers, while the existence of such responsibility cannot be contractually waived, there is a possibility of assigning partial responsibility according for distinct stages of data processing. Thus, different degrees of responsibility can be assigned proportionately to the participation of the respective data controller to the data processing. The same can be derived from most recent and previously established case law from the CJUE (Case C-40/17 Fashion ID).

Against this legal background, the participating nodes within a decentralized identity infrastructure could qualify as joint data controllers for the transactional data that they to verify, store, and put on/off chain. Even if the means and purposes of the data processing and the architectural design rules that will govern the safe and secure data processing are decided in a less decentralized manner by a single entity, the participation on the network can lead to such a qualification especially given the progressively expansive case law on the responsibility allocation of data controllers. However, each joint controller can only be considered responsible within the limits of the data processing they are facilitating. The more pragmatic approach -adopted by the responsible bodies and case law- in the determination of liable actors and the allocation of liability between them signifies that actors can be qualified as joints controllers when they exert "a decisive influence over the collection and transmission" of the personal data, without necessarily having access to the data in question and where there is joint determination of the purposes of the processing.

The architectures of decentralized identity could also lead to the qualification of the data subjects as data controllers with regards to their own data [9], which is a legal concept that has not yet been tested in court but which has been indirectly suggested through developed case law.

4 Conclusion

Decentralized -or self-sovereign- identity is an emerging concept that should be regarded critically for its purported benefits in providing solutions for issues like private and secure exchange of personal data among actors that do not necessarily trust each other and without the mediation of an institution acting as the certifying authority.

Whether it consists of a bottom-up approach to establish community-driven norms and solutions to the systemic problem of data-intensive technologies, or company investments in developing a product that corresponds to similar societal needs, or even a public institutions aiming to provide innovative solutions for its citizens, compliance with data protection norms is key. Legal compliance can be the gateway for a lot of these projects to reach some level of recognition and usability but also it can be the tool that ensures that these projects deliver on their promise to redesign personal data exchanges. In that regard, the GDPR is a malleable enough framework to convey both fundamental protections necessary to data protection but also to accommodate a decentralized network of actors that deploy technological architectures in order to achieve a high level of security and privacy.

References

1. Finck M.: Blockchain regulation and governance in Europe, Cambridge University Press, United Kingdom (2020).
2. Halpin H, Decentralizing the Social Web. Can Blockchains Solve Ten Years of Standardization Failure of the Social Web? In S. S. Bodrunova et al. (Eds.), INSCI 2018 Workshops, LNCS 11551, Springer, (2019).
3. Sullivan C., Burger E., "E-Residency and Blockchain" (2017) 33 Computer Law & Security Review 460, 475.
4. Wang F, De Filippi P Self-Sovereign Identity in a Globalized World: Credentials-Based Identity Systems as a Driver for Economic Inclusion, (2020), Frontiers in Blockchain, DOI 10.3389/fbloc.2019.00028
5. Giannopoulou A, Algorithmic systems: the consent is in the details?, Internet Policy Review, 2020, forthcoming.
6. Bodo B, Giannopoulou A, The Logics of Technology Decentralization: the Case of Distributed Ledger Technologies. In M. Ragnedda, & G. Destefanis (Eds.), Blockchain and Web 3.0: Social, Economic, and Technological Challenges Routledge, 2020.
7. Gooddell G, Tomaso A, A Decentralized Digital Identity Architecture, (2020), Frontiers in Blockchain, DOI 10.3389/fbloc.2019.00017
8. Finck M, Pallas F (2019) They Who Must Not Be Identified - Distinguishing Personal from Non-Personal Data Under the GDPR. International Data Privacy Law, 2020; Max Planck Institute for Innovation & Competition Research Paper No. 19-14.
9. Edwards L, Finck M, Veale M, Zingales N, Data subjects as data controllers: a Fashion(able) concept?, Internet Policy Review, 2019.

Acknowledgments

The Lab has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 759681.